Project Alternus Vera: Detecting Disinformation

“Alternative Truths”? Nonsense ☺!

Dr. Ali Arsanjani

San Jose State Univ

Your objective in ~~preventing~~ mitigating disinformation

In WhatsApp's "Tips to help prevent the spread of rumors and fake news," three of the seven tips focus on preventing the spread of fake news.

One of the biggest problems with fake news is not necessarily that it gets written, but rather that it gets spread.

Therefore, detecting it is paramount

disinformation

The parliamentary committee on digital, culture, media and sport in the United Kingdom recommended against using "fake news" in favour of more specific terms:

"The term 'fake news' is bandied around with no clear idea of what it means, or agreed definition.

The term has taken on a variety of meanings, including a description of any statement that is not liked or agreed with by the reader.

*We recommend that the Government rejects the term 'fake news,' and instead puts forward an agreed definition of the words '**misinformation**' and '**disinformation**.'"*

Disinformation is the deliberate creation and/or sharing of false information in order to mislead.

Misinformation is the act of sharing information without realizing it's wrong.

the kinds of fake content you're likely to see online:

- **Fabricated content**: completely false content.
- **Manipulated content**: content that includes distortions of genuine information or imagery — a headline, for example, that is made more sensationalist to serve as "clickbait."
- **Imposter content**: material involving impersonation of genuine sources — by using the branding of an established news agency, for instance.
- **Misleading content**: information presented in a misleading way — by, for example, presenting comment as fact.
- **False context of connection**: factually accurate content that is shared with false contextual information — for example, a headline that does not reflect the content of an article.
- **Satire and parody**: humorous but false stories presented as if they are true. Although this isn't usually categorized as fake news, it may unintentionally fool readers.

First Draft News, an organization dedicated to improving skills and standards in the reporting and sharing of online information, explains the fake news environment and proposes 7 types of fake content:

1. False Connection: Headlines, visuals or captions don't support the content
2. False Context: Genuine content is shared with false contextual information
3. Manipulated content: Genuine information or imagery is manipulated
4. Satire or Parody: No intention to cause harm but potential to fool
5. Misleading Content: Misleading use of information to frame an issue/individual
6. Imposter Content: Impersonation of genuine sources
7. Fabricated content: New content that is 100% false

https://firstdraftnews.com/fake-news-complicated/

FIRSTDRAFT

7 TYPES OF MIS- AND DISINFORMATION

DataSets

1. BuzzFeedNews: This dataset comprises a complete sample of news published in Facebook from 9 news agencies over a week close to the 2016 U.S. election from September 19 to 23 and September 26 and 27. Every post and the linked article were fact-checked claim-by-claim by 5 BuzzFeed journalists. It contains 1,627 articles 826 mainstream, 356 left-wing, and 545 right-wing articles.
2. LIAR: This dataset is collected from fact checking website PolitiFact. It has 40 K human labeled short statements collected from PolitiFact and the statements are labeled into six categories ranging from completely false to completely true as pants on fire, false, barely-true, halftrue, mostly true, and true.
3. BS Detector: This dataset is collected from a browser extension called BS detector developed for checking news veracity. It searches all links on a given web page for references to unreliable sources by checking against a manually compiled list of domains. The labels are the outputs of the BS detector, rather than human annotators.
4. CREDBANK: This is a large-scale crowd-sourced dataset of around 60 million tweets that cover 96 days starting from Oct. 2015. The tweets are related to over 1,000 news events. Each event is assessed for credibilities by 30 annotators from Amazon Mechanical Turk.
5. BuzzFace: This dataset is collected by extending the BuzzFeed dataset with comments related to news articles on Facebook. The dataset contains 2263 news articles and 1.6 million comments discussing news content.
6. FacebookHoax: This dataset comprises information related to posts from the facebook pages related to scientific news (non- hoax) and conspiracy pages (hoax) collected using Facebook Graph API. The dataset contains 15,500 posts from 32 pages (14 conspiracy and 18 scientific) with more than 2,300,000 likes.
7. FakeNewsNet: This dataset comprises fake and real news pieces collected from fact-checking websites PolitiFact and GossipCop. It contains news articles content, tweets related to news articles and their social engagements including replies, retweets, and favorites. In total dataset contains nearly 2 million tweets related to fake and real news pieces along with their engagements and user profiles of users interacted with these tweets.

The Liar Liar Data Set

label	statement	topics	person	job	location	affiliation
barely-true	We have less Americans working now than in the 70s.	economy,jobs	vicky-hartzler	U.S. Repres	Missouri	republican
pants-fire	When Obama was sworn into office, he DID NOT use the Holy Bible, but inste	obama-birth-certificate,r	chain-email			none
FALSE	Says Having organizations parading as being social welfare organizations and	campaign-finance,congr	earl-blumenauer	U.S. represe	Oregon	democrat
half-true	Says nearly half of Oregons children are poor.	poverty	jim-francesconi	Member of	Oregon	none
half-true	On attacks by Republicans that various programs in the economic stimulus pla	economy,stimulus	barack-obama	President	Illinois	democrat
FALSE	Says when armed civilians stop mass shootings with guns, an average of 2.5 p	guns	jim-rubens	Small busin	New Hamp	republican
TRUE	Says Tennessee is providing millions of dollars to virtual school company for r	education,state-budget	andy-berke	Lawyer and	Tennessee	democrat
FALSE	The health care reform plan would set limits similar to the socialized system	health-care	club-growth			none
TRUE	Says Donald Trump started his career back in 1973 being sued by the Justice D	candidates-biography,div	hillary-clinton	Presidential	New York	democrat
half-true	Bill White has a long history of trying to limit or even disenfranchise military v	military	republican-party-te	xas	Texas	republican
half-true	John McCains chief economic adviser during the 08 race estimated that Trump	economy	tim-kaine	U.S. Senato	Virginia	democrat
FALSE	Says 21,000 Wisconsin residents got jobs in 2011, but 18,000 of them were in	job-accomplishments,job	kathleen-vinehout			democrat
half-true	State revenue projections have missed the mark month after month.	state-budget	steve-henson	State Senat	Georgia	democrat
TRUE	The median income of a middle class family went down $2,100 from 2001 to	income,new-hampshire-	joe-biden	U.S. senato	Delaware	democrat
barely-true	Every citizen is entitled to the freedom of speech, but no one should have the	gays-and-lesbians	david-dewhurst	Lieutenant g	Texas	republican
half-true	Rick Perry has advocated abandoning Social Security, scuttling Medicaid and e	medicaid,social-security,	margaret-carlson	Columnist	District of	none
half-true	Two thirds to three quarters of people without [health] insurance in Rhode Isl	health-care,poverty,publ	elizabeth-roberts	Lieutenant (	Rhode Isla	democrat
mostly-true	Congress has spent 66 of the first 100 days of this term in recess.	congress	john-barrow	Congressma	Georgia	democrat
barely-true	Mark Sharpe has lowered property taxes by 17 percent.	candidates-biography,tax	mark-sharpe	Hillsboroug	Florida	republican
pants-fire	Says Iowa Gov. Terry Branstad chartered a plane to remove 124 young illegal	immigration	chain-email			none
half-true	If you dont buy cigarettes at your local supermarket, your grocery bill wont go	taxes	philadelphia-daily-n	ews		none
pants-fire	Says President Barack Obama has said that everybody should hate the police.	civil-rights,crime,crimina	rudy-giuliani	Attorney	New York	republican
TRUE	Georgia has had ∫more bank failures than any other state. ∫	bankruptcy	lynn-westmoreland			republican
pants-fire	Bank of America could create 878,300 jobs with benefits if they spent their 20	financial-regulation,jobs,	facebook-posts	Social media	posting	none
half-true	Thom Tillis cut almost $500 million from education.	children,congress,econor	emilys-list			organization
barely-true	If people work and make more money, they lose more in benefits than they w	poverty,welfare	marco-rubio	U.S. Senato	Florida	republican
mostly-true	We are poised to get rid of over 1,000 more regulations in 2012.	government-regulation,r	rick-scott	Governor	Florida	republican
pants-fire	A flight from Atlanta to Houston was canceled due to a terrorist dry run.	crime,terrorism	facebook-posts	Social media	posting	none

The Liar Liar Data Set

barely tru	FALSE	half true	mostly tr	pants on	context (venue / location)
1	0	1	0	0	an interview with ABC17 News
11	43	8	5	105	
0	1	1	1	0	a U.S. Ways and Means hearing
0	1	1	1	0	an opinion article
70	71	160	163	9	interview with CBS News
1	1	0	1	0	in an interview at gun shop in Hudson, N.H.
0	0	0	0	0	a letter to state Senate education committee chairwom
4	5	4	2	0	a TV ad
40	29	69	76	7	the first presidential debate
3	1	1	3	1	an e-mail
8	3	15	15	0	a speech at the Democratic National Convention in Phil
1	1	1	1	0	remarks
0	0	1	0	0	a press release
11	10	21	16	4	speaking at New Hampshire,Äôs Plymouth State Univer
8	8	10	5	5	a press release
0	0	1	0	0	a politics column.
1	0	2	0	0	a panel discussion on "A Lively Experiment"
0	0	1	1	0	a letter
1	0	0	0	0	a campaign mailer
11	43	8	5	105	a chain email
0	0	1	0	0	In an editorial
9	11	10	7	3	an interview on Fox News
1	1	3	0	0	a meeting
14	18	15	11	36	a Facebook post
1	0	3	2	0	a television ad
33	24	32	35	5	his book, "American Dreams"
28	23	38	34	7	speech at CPAC
14	18	15	11	36	a post on Facebook

36 Factors of Veracity and Misinformation

NodeRank	Political affiliation	Sentiment Analysis	Topics	Post/Social media activities
Spam	Visual based	Sensationalism	Writing Style	Reliable source
Context Veracity	Verifiable Authenticity	Misleading intentions	Confirmation Bias	Psychology utility
Content Statistics	Social credibility	Frequency heuristic	Credibility and Reliability	Stance Detection
Source Reputation	Echo Chamber	Location / Geography	Education	Biases
News Coverage	Malicious Account	Naive Realism	Network-based	Corpus Structure
Political bias	Event Coverage	Title vs Body	ClickBait	Neural [Micro-patterns of] Misinformation
Stance Detection	Toxicity	BERT - Transformers		

Fake News Detection on Social Media: A Data Mining Perspective: https://arxiv.org/pdf/1708.01967.pdf

1. News outlets,
2. social media,
3. aggregators,
4. search

Choose a subset

36 Factors

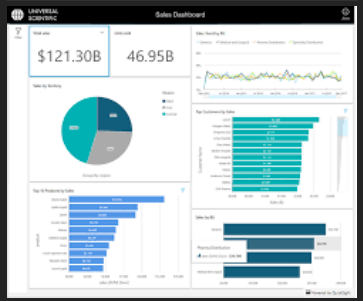
Metrics

e.g., credibility

e.g., Hate speech

Visualization of Veracity Vectors for:

1. Individual
2. Comparative,
3. Aggregate,
4. Veracity Tensors



	AV	Vectors	Veracity
cnn			.85
fox			.45
npr			.80
twt			.20

intuition

News

Dataset1-3

1. BuzzFeedNews: Th
2016 U.S. election f
by 5 BuzzFeed jour
2. LIAR: This dataset i
and the statements
halftrue, mostly true
3. BS Detector: This d
links on a given wet
outputs of the BS de
4. CREDBANK: This i
tweets are related t
5. BuzzFace: This dat
contains 2263 news
6. FacebookHoax: Thi
conspiracy pages (
scientific) with more
7. FakeNewsNet: This
contains news articl
total dataset contair

accuracy : .45

Factor 1: Model

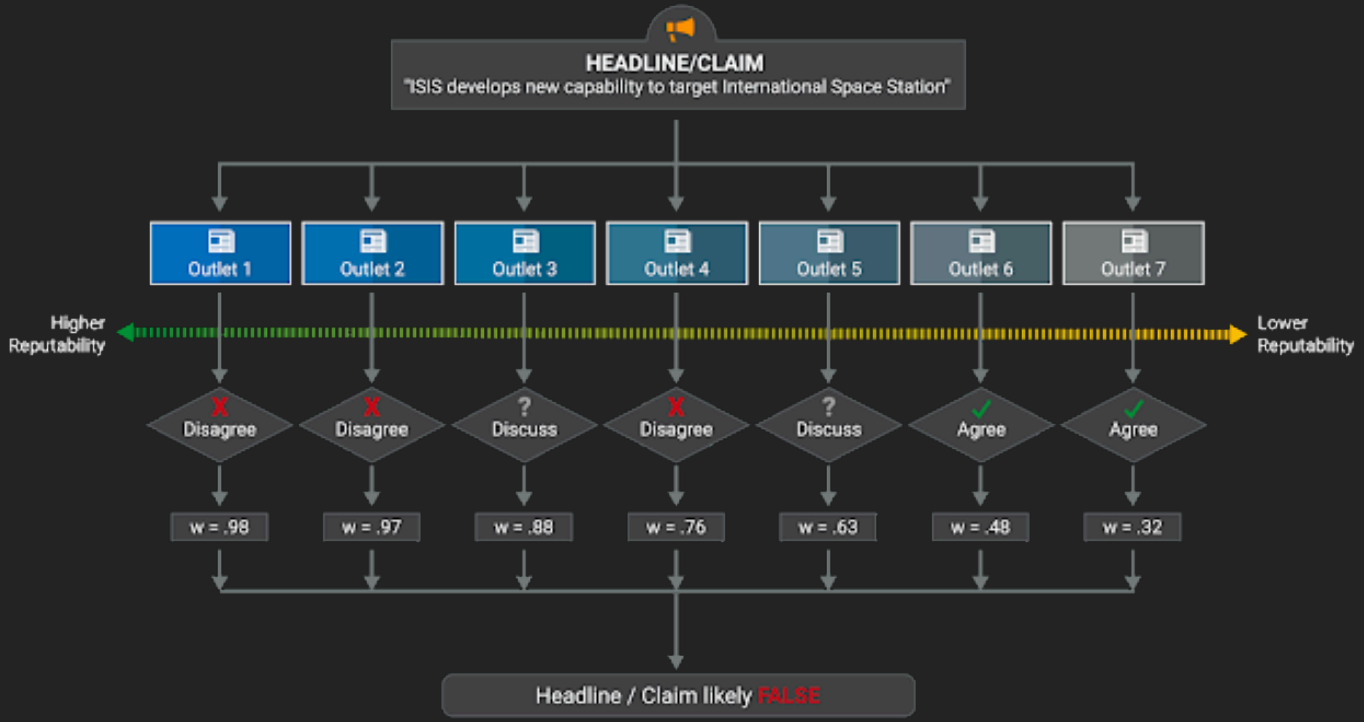
Factor n: Model

accuracy : .55

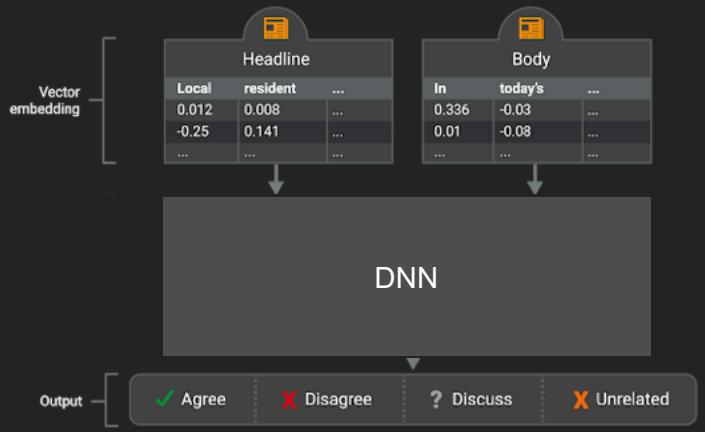
Polynomial: A1x1+ a2x2

Score : .76

Sample Factor: Stance Detection

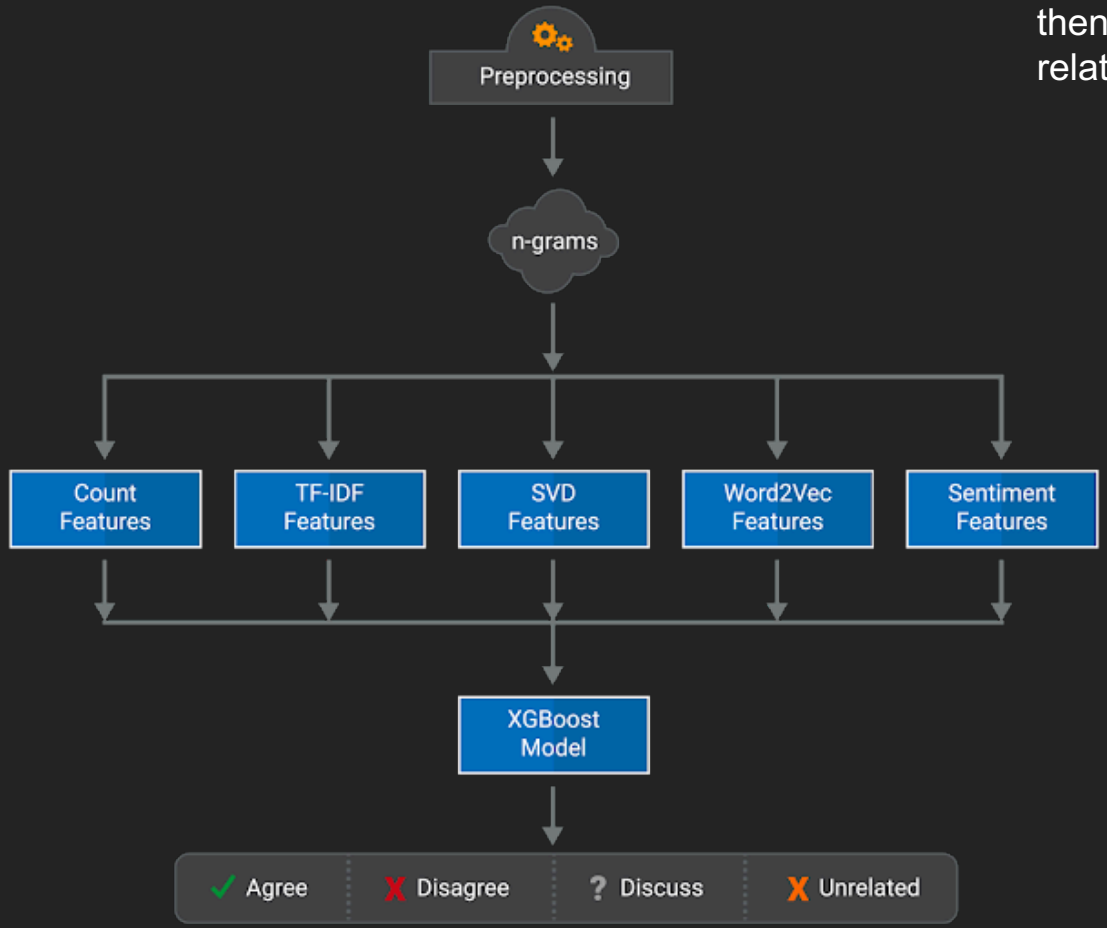


Sample Factor: Stance Detection



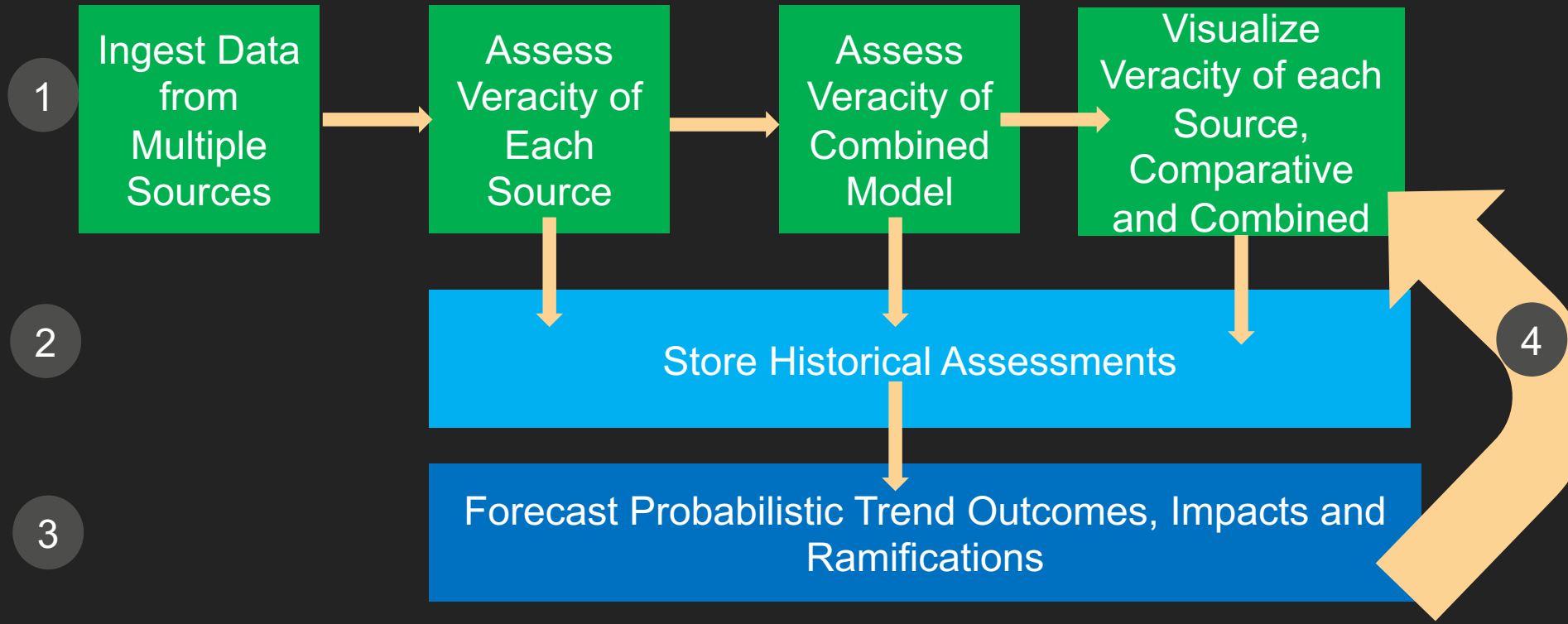
Sample Factor: Stance Detection

This model inputs other text-based features derived from the headline and body of an article, which are then fed into Gradient Boosted Trees to predict the relation between the headline and the body.



•After exploring the dataset, a few features that are likely to be informative of headline/body relationships became obvious -- for example:

The number overlapping words between the headline and body text;
•Similarities measured between the word count, 2-grams and 3-grams; and
•Similarities measured after transforming these counts with term frequency-inverse document frequency weighting and Singular Value Decomposition


1
Ingest Data from Multiple Sources
Assess Veracity of Each Source
Assess Veracity of Combined Model
Visualize Veracity of each Source, Comparative and Combined
2
Store Historical Assessments
3
Forecast Probabilistic Trend Outcomes, Impacts and Ramifications
4

Software Engineering for ML

DrArsanjani

Week 9 Supplement

AV Project: Colab for Code

Class for each Factor

Class Factor()

- Prep
  - Def Read the csv(s) : (def a func!!)
    - Access the common googledrive
    - LL
    - FNC
    - Your data set for your factor
  - Data Preparation depends on the algorithm
    - Some algo with numbers
    - Some work with text/unstru (MNBayes)
    - BERT : string → delimiters in the string
    - Categ → one hot encode
    - Numeric → leave them!
    - LDA
  - Algorithm Select 4-5 algorithms to run
  - Metrics ? To measure model perf
    - F1, recall, precision, accuracy, confusion matrix (print out)

Alg Comparisons based on Metrics

	MNB	XGB	LR	DNN
accuracy				
prec				
Recall				
f1				

You can run individually or use a PIPELINE!!!

Save and Load your models and big imports

- Train
  - Pickle the model (dump), consider tar.gz.
    - You may want to tar gz if cloud deployment
- Inference, Serving, Prediction
  - Pickle: read (load)
  - myFeature = pickle.load(“factor_01_word_freq”)
  - myFeature.predict(article)

Data : GoogleDrive

MLFall2020

- <Team name folder> Avengers
  - LL
  - <Factor –> Misleading Intentions
    - “Dataset”
    - “Model”
      - » Avengers_Factor_TitleVsBody.pkl
      - » Avengers_Factor_ LDA Topic Features
      - » Avengers_Factor_ Text Rank
      - » Avengers_Factor_ Misleading Intentions